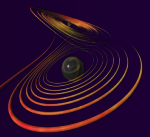
OpenAFS

Das weltweite Filesystem

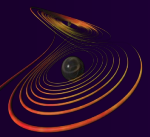
Gliederung

- Übersicht
- Globale Sicht - logisch
- Lokale Sicht - phsikalisch
- Filesystem - logisch
- Vorteile/Nachteile
- Demos
- Fragen?



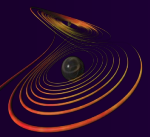
Einführung

- Entwickelt in der Carnergie Mellon University als “Andrew File System”
- Von Transarc Corp. kommerziell vertrieben
- IBM übernahm Transarc und somit AFS
- Ab dem Jahr 2000 Open Source unter der IBM Public Licence
- Seitdem aktive Fortentwicklung durch viele Freiwillige und sponsored Personen
- http://www.openafs.org



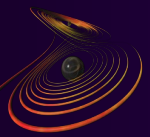
Eigenschaften

- IP basiertes verteiltes Netzwerkfilesystem
- Server-Client Struktur
- Daten werden in Volumes verwaltet
- Read/Write & Read only Volumes
- ACL für Verzeichnisse
- Gruppenverwaltung
- Lokaler Cache auf Client-PC
- Transparente Administration



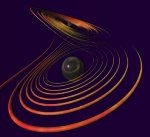
Plattformen

- OpenAFS ist derzeit erhältlich für:
- Windows 2000,XP,2003,Vista
- Linux (Debian, RedHat, SuSe,...)
- Mac OS X (10.4 - Tiger)
- Digital UNIX 4.0d /Tru64
- NetBSD/OpenBSD/FreeBSD
- Solaris 9
- HP-UX 11i
- IRIX
- AIX



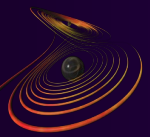
Server-Client Topologie

- Eindeutige Zuordnung der Server zu einer “Zelle”, die eine Einheit darstellt
- Root des AFS Filesystemes ist /afs/, dort werden die Zellen direkt gemountet
- Verwaltung der Zellen für den Client in einer “CellServDB” Textdatei oder aus den entsprechenden Nameservern der Zellen



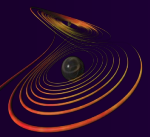
Zellenstruktur

- Je Zelle min. 1 Database Server
- Je Zelle min. 1 Fileserver
- Clients
- Kommunikation via TCP/IP
- Somit ortsungebunden



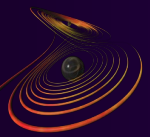
Database Server

- Der Database Server besteht aus mehreren Instanzen:
  - Ptserver - User & Gruppen Datenbank
  - Vlserver - Volume Datenbank
  - BOS Server - Master Controller
- 3 Database Server empfohlen bei großen Zellen
- Beim Startup wird ein Master Server auserkoren und die Database Server gleichen ihre Configs ab
- 2 stündlich propagation Config an File/DatabaseServer
- Mind. Einer muß erreichbar sein



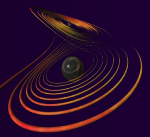
Fileserver

- Besteht aus folgenden Prozessen:
  - Fs server – eigentlicher Fileserver
  - Salvager – Partition salvager
- 2++ Fileserver empfohlen
- Beim startup startet der salvager automatisch und überprüft die Daten auf den Partitionen, sonst explizit per Befehl
- Jeder Fileserver sollte erreichbar sein



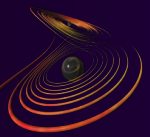
Lokales Filesystem Server

- Database wir ins System gespeichert (/var/lib/)
- Daten wird auf eigenen Partitionen gespeichert
  - /vicepa, /vicepb, /vicepaa,...
  - Kann auch symlink sein
  - Muß beim startup vorhanden sein!
  - Journal FS geht auch, ist nicht empfohlen
  - Eigenes Dateiformat, nicht plain (perl script zur Datenextraktion)
  - Inode (Solaris)/Namei Datenstruktur



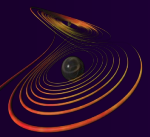
Client

- Besteht aus folgenden Komponenten:
  - Cache (RAM oder Disk)
  - Anbindung ans lokale System
    - Z.B. Windows per smb/ UNC Pfad
    - Linux direkt als Mountpunkt /afs/
  - Diverse kleine tools zum Verwalten der ACLs, Volumes,....
- Muß/sollte Kontakt zu Fileserver und Database Server haben



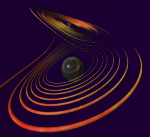
Lokales Filesystem Client

- Cache Verzeichnis/Datei
  - Windows 100 MB C:\win\tmp\AFSCache
  - Linux: /var/cache/openafs variable Grösse
- Config in /etc/openafs
- Linux: Kernel Module für Client (Server ohne)
- Unter Windows Laufwerk auf \\AFS\Zelle mounten!



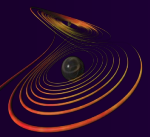
Die Volumes

- Logisch ist das FS in Volumes auf den Fileservern strukturiert (ähnlich wie LVM)
- Selbst /afs/ ist ein Volume (root.afs)
- Ein Volume enthält die Daten, ACLs und Quota
- Es gibt von jedem Volume EINE Read-Write Instanz, ggf. EINE Backup Instanz und Readonly Instanzen
- Volumes sind mehrfach frei einhängbar
- Volumes sind transparent und frei auf den Fileservern und Partitionen verschiebbar



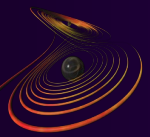
Cache

- beim Lesen & Schreiben Cache in der Pipe
- Chunk-Cache, KEINE ganzen Dateien
- Max. 2 GB unter 32bit Win, unter Linux 4GB sinnig
- Konfigurierbar (chunkgröße, Statuscache,..)
- Beim lesen wird erst Cache gefragt, bei miss der Fileserver, bei hit wird Aktualität beim Server kontrolliert
- Schreiben erst in Cache, dann auf fileserver, Probleme mit WinSCP & großen Dateien



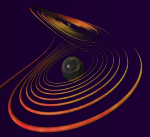
Die Volumes

- Logisch ist das FS in Volumes auf den Fileservern strukturiert (ähnlich wie LVM)
- Selbst /afs/ ist ein Volume (root.afs)
- Ein Volume enthält die Daten, ACLs und Quota
- Es gibt von jedem Volume EINE Read-Write Instanz, ggf. EINE Backup Instanz und Readonly Instanzen
- Volumes sind mehrfach frei einhängbar
- Volumes sind transparent und frei auf den Fileservern und Partitionen verschiebbar



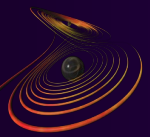
Das Filesystem

- Root ist /afs/
- De-facto Standard ist:
  - /afs/zellen-resolve/ RO eingehangen
  - /afs/.zellen-resolve/ RW eingehangen
  - Home-Verzeichnisse unter ../home/
- Somit alle Zellen mit cd /afs/... erreichbar sobald entsprechendes Volume eingehangen (oder fakeroot Option aktiviert)



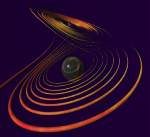
Die Userverwaltung

- Mit dem „pts“ Befehl werden User/Gruppen verwaltet
- Es gibt spezielle Gruppen:
  - system:anyuser – JEDER AFS Client
  - system:authuser – jeder IN der Zelle authentifizierte User
  - system:adminstrators – Zellen Admins
  - system:backup – Backup User
  - system:ptsviewers – Gruppenverwaltung, protection database Einsicht



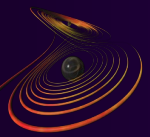
Userverwaltung 2

- User anlegen, Gruppe erzeugen
- Gruppen haben negative ID, (frei) wählbar
- User haben positive Ids (auf Linux ID gemapped)
- Gruppen sind Zellenweit
- User können 20 private Gruppen erzeugen
- „spezielle“ User sind IP – der PC mit der IP ist dann entsprechend der User (das System auf dem PC hat dann die Rechte, NICHT nur der User!)
- User werden alle 2h propagandiert, Gruppen schneller



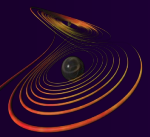
Zellenstruktur

- Volumes zu erzeugen:
  - User – homeverzeichnisse
  - Bin – binaries
  - Data – Daten
  - Www – Webseite
  - Ftp – FTP Daten
  - Win – win Profiles
  - Archive – Archive
  - Demo - Demos



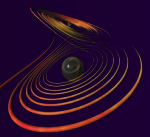
Verzeichnisbaum aufbauen

- Volumes mounten:
  - Vos create server partition volumename
  - User RW mounten als /afs/zelle/home
  - Ftp RO mounten als /afs/zelle/ftp
  - Archive RO mounten als /afs/zelle/archive
- Den gemounteten Volumes Quota geben
  - Fs setquota pfad 1000
- Den erstellen Verzeichnissen ACLs zuweisen
  - Fs setacl pfad User/Gruppe ACL
- RO Kopie des Volumes erstellen



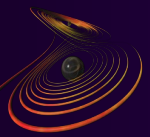
Die ACLs

- Für JEDES Verzeichnis gelten folgende ACLs:
  - L – lookup – Dir Inhalte einsehen
  - R – read – Dateiinhalte lesen
  - I – insert – Dateien anlegen
  - W – write – Dateien beschreiben
  - D – delete – Dateien löschen
  - K – lock – Dateien locken
  - A – administer – ACLs verwalten



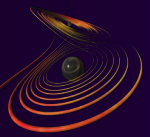
Verzeichnisse, Volumes,..

- ACLs können auf sämtliche Verzeichnisse gesetzt werden
- ACLs werden vererbt in neu erstellte Verzeichnisse, jedoch NICHT auf gemountete Volumes!
- Quota gilt NUR auf Volumes!
- Wird ein Volume doppelt gemountet, gelten an beiden Pfaden die selben Quota/ACLs
- Geschickte Vergabe sorgt für automatische Zugriffsverteilung



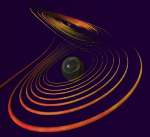
Besondere Pfade

- Es existiert der Special Pfad @sys :
- /afs/zelle/bin/@sys/ wird aufgelöst nach dem Sysnamen vom OpenAFS Clienten (fs sysname):
  - i686_linux26
  - i386_linux24
  - tru64
  - ppx_linux24
  - Win32
  - Win64
  - .....



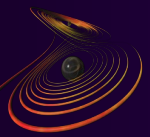
@sys

- Dieser Pfad kann z.B. für Architekturspezifische Software genutzt werden:
  - /afs/zelle/bin/@sys/xchat/xchat lädt automatisch das passende binary auf linux,win, mac,...
- In Verbindung mit dem IP User hat ein Multi-Boot PC so IMMER ein Programm unter dem selben Pfad erreichbar
- Sogar MS Office funktioniert aus dem OpenAFS heraus



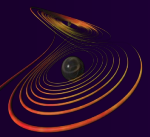
Nachteile

- Steile Lernkurve
- Netzwerk muss da sein fürs Schreiben/Server
- Nur ca. 60.000 Dateien/Verzeichnis
- Speichert nur den Datenstream (Windows, NTFS)
- Auf Netzgeschwindigkeit begrenzt
- Client mit Kernelmodul auf Linux
- Backup nur volle Volumes oder spezial Software (ACLs, Quota, Mountpoints....)



Vorteile

- Netzwerkdateisystem – weltweit Zugriff
- ACLs und Quota integriert
- Einfach administrierbar, verteilt administrierbar
- Lastverteilung bei RO Volumes
- Volumeaktionen ohne Unterbrechung des Betriebs
- Universell
- Skalierbar auf >100.000 User
- Integrierte rudimentäre Backuplösung



Für Wen?

- viel Aufwand für Privatperson mit 2 PCs
- Lohnt ab 2 Standorten (ohne/mit) VPN
- Ideal bei grösseren Installationen:
  - Universitäten
  - Internationale Firmen/Konzerne
  - Verwaltungen
  - Linuxgruppen ;-)